Finding Records in Social Media:
A Natural Language Processing Fundamentals Exploration

Babatunde Kazeem Oladejo [1][0000-0002-3879-6345] *, Sunčica Hadžidedić [2][0000-0001-9026-8737] and Emir Ganić [3]

[1,3] Department of Computer Science and Information Systems, University Sarajevo School of Science and Technology, 71210 Ilidza, Bosnia and Herzegovina
[2] Department of Computer Science, Durham University, South Road, Durham, DH1 3LE, United Kingdom
[1] babatunde.oladejo@stu.ssst.edu.ba, +387 60 326 3039
[2] suncica.hadzidedic@durham.ac.uk, +44 (0) 191 3347346
[3] emir.ganic@ssst.edu.ba, +387 33 975 034

* Corresponding author

Abstract

Social media postings are now routinely used as proof of activities, events, or transactions in news media, academic institutions, governments, judicial courts, commerce, and various other organisations. The need to preserve social media content as records has drawn the interest of academic researchers, industry professionals, and policy makers. Despite the importance of this research area, selection of records from a pool of social media content remains an area of low research activity. This paper explores the use of Natural Language Processing methods to classify and select records from a pool of tweets (twitter social media content). We experiment with various characteristics of the data and NLP parameters with the goal of determining optimal parameters for training a supervised machine learning classifier. This paper can serve as an aid for understanding the fundamental elements of automating the selection of social media records.

Keywords: Social Media Record, Record Selection, Record Classification, Machine Learning, NLP

1.0 Introduction

When Darnella Frazier, a teenager from Minneapolis, MN, USA decided to post a video of George Floyd's police brutality on her Facebook social media page, she did not realize how big the impact would be. Her attorney, Seth Cobin said in the Star Tribute [1] "She had no idea she would witness and document one of the most important and high-profile police murders in American history". The social media post went viral, bypassing mass media, crime reporting agencies and all traditional record capture to become the central evidence against the police officers charged in the court case.

Beyond the law and order example above, the use of records from social media has become prevalent in education [2], mass media [3], medical practice [4], government administration [5], corporate business [6], and other human-cultural organizations. This engaged utilization of records from social media prompts the questions: what is a social media record? What are the criteria for the selection of social media records from a pool of social media content?

What is a social media record?

The US National Archives and Records Administration (NARA) in its Bulletin 2014-02 [7] defined social media records as social media content that fulfils the record criteria of material that is recorded, made or received in the course of official business, regardless of its form or characteristics, and is worthy of preservation. A social media record must have content, context, and structure along with associated metadata (e.g., author, date of creation), and be properly maintained to ensure reliability and authenticity. To assist agencies in determining the record status of social media content, NARA further specified a (non-exhaustive) list of qualification questions:

- Does it contain evidence of an agency's policies, business, or mission?
- Is the information only available on the social media site?
- Does the agency use the tool to convey official agency information?
- Is there a business need for the information?

A 'yes' answer to any of the questions would make the social media content selectable as a record.

Although developed for US federal agencies, the NARA social media record policy addresses several of the issues and challenges facing other organizations [8]. An establishment's record policy usually dictates its record

selection and records classification agenda. Records selection and records classification are fundamental processes of Records Management [9] and are essential to any study in a new and developing sub-section, such as social media. This paper will focus on these two interwoven elements in our exploration of social media records management.

Finding social media records in a pool of social media content is a challenging problem to solve because, practically, any content can be a record [10]. For example, a casual "hello" social contact between an investigator and the criminal being investigated might become a record if both parties claim not to know each other. However, such a casual "hello" social contact between ordinary people would not be considered records.

For this research, instead of attempting to directly decipher social media records from a pool of social media content, we took the approach of using an already curated record source and subsequently examine the characteristics of the records. A successful exploration of the data would enable us to use Natural Language Processing (NLP) methods to elicit features from the textual content of the records. These features could then be used to train a machine learning classifier to identify new records in a new pool of social media content. Our chosen dataset is twitter social media postings in the news publications of the British Broadcasting Corporation (BBC) and the New York Times (NYT). Articles published by news publishers are often referred to as credible sources of information curated for the public under due editorial processes, and therefore acceptable as records [11].

With the foregoing approach to the identification of the social media record and its management, we articulate research goals as:

1) Explore the selected social media record dataset for its properties, which can assist in the implementation of an automated record selection initiative.
2) Explore the use of NLP techniques to classify records in the selected social media dataset and determine the best set of parameters and algorithms for the training of a machine learning classifier.

The rest of this paper is organized as follows. Section 2 discusses the literature related to the research objectives. Section 3 covers data collection and characteristics of the dataset, while Section 4 details the pre-processing work done on the dataset. In Section 5, we annotate the experiments performed and discuss the results. Section 6 summarizes the research and discuss potential areas for future improvement and limitations.

2.0 Related Works

Our search of academic databases namely, Google Scholar, DeepDyve and Springer Link for articles related to Social Media Records Management returned only about 200 articles. Several of the articles belong to the archivist, records management and law domains and only describe the nature of the social media record from theoretical and application practice viewpoints [12], [13], [14]. We found fewer articles that describe the computing nature of social media records, which is the primary focus of our research. Van Wyk and Starbird [15] used the term social media record to describe a collection of messages in a time-framed event such as an earthquake. The idea of the social media record being a collection of related messages was also used by Liu et. al [16], but with a patient's social media message collection as the record. The researchers created a custom algorithm called SocInf to evaluate their hypothesis of Membership Inference attacks by potential hackers. SocInf's performance was compared against three other machine learning models trained on logistics regression, Xgboost and BigML, a cloud-based platform. SocInf was found to be the best performer in their experiments.

Given the lack of computing research materials on social media records management, we considered studies on social media topic classification that use NLP methods and approaches in line with our research goals. Topic classification deals with the grouping of social media content [17] and is well aligned with record classification, where record is a subset of content [7]. We narrowed down the vast number of search results to articles that address news related topical classifications of social media content in line with our research dataset and experiment objectives and report their findings in the following text.

Dataset Features and Pre-processing

Iman, Zahra, et al. [17] conducted a large longitudinal study of twitter topic classification of using over 800 million English language tweets between 2013 and 2014. The study found that Naïve Bayes is an effective topical learner which could generalize and generate new previously unseen news-worthy topics after the year-long training. The authors compared the effectiveness of various features and found hashtags, mentions, locations to be amongst the best features for training their classifier.

Perreault and Ruths [18] found that manual labelling for supervised topic classification result in higher accuracy and precision than unsupervised methods, but labelling can be labour intensive, time-consuming, and expensive. Semi-automated labelling can also achieve good results given adequate use of training instances [19]. To improve the quality of social media content for classification, pre-processing is essential. Pre-processing, textual clean-up or text massaging

tasks often include removal of stop-words, part of speech (POS) tagging and replacement of non-standard words (such as “lmao”, “cuz”, “lol”, etc. with the standard equivalents) [20].

NLP Algorithms and Performance Measurements

Working with a team of journalists to identify newsworthy events that were likely to become rumours, Zubiaga et al. [21] used a linear-chain Conditional Random Fields (CRF) algorithm to learn the dynamics of information during breaking news and classified the information as rumour or non-rumour. A performance comparison between CRF and other classifiers (SVM, Random Forest, Naïve Bayes, and Maximum Entropy) was conducted. While SVM best exploited the social (twitter) features, CRF was better with source (news publisher) features. The overall result showed that CRF performed best in terms of precision but lacked behind in recall. The study concluded that when all constraints were considered, CRF outperformed the other classifiers in the detection of rumours.

A similar comparative evaluation of ML algorithms for the detection of credible news was conducted by Hassan et. al. [22]. The researchers compared 5 algorithms, namely, Linear Support Vector Machines (LSVM), Logistic Regression (LR), Random Forests (RF), Naïve Bayes (NB) and K-Nearest Neighbours (KNN). They found that the best performance was achieved with LSVM using a combination of unigrams and bigrams as features prioritized by TF-IDF (Term Frequency – Inverse Document Frequency).

The challenge of securing adequate data for supervised machine learning of fake news was addressed by Helmstetter and Paulheim [23] by using a technique called weakly supervised learning. A dataset of tweets was automatically labelled by the source reliability, i.e. trustworthy or untrustworthy source, and a classifier trained on the dataset. The classifier was then repurposed for a different classification target, i.e., the classification of fake and non-fake tweets. Interestingly, the labels were not always accurate according to the new classification target (i.e. not all tweets by an untrustworthy source turned out to be fake news, and vice versa), the research show that despite the inaccuracy of the original dataset, fake news could be detected with an F1 score of up to 0.9 using the XGBoost classifier.

Salminen, Joni, et al. [24] created a detailed taxonomy of online hate types and people targeted as part of an effort to automate the detection of online hate expressions. The researcher created ML models that classifies the hateful comments, experimenting with Logistic Regression, Decision Tree, Random Forest, Adaboost, and SVM. The study found that SVM performed the best for the dataset, with an average F1 score of 0.79.

Overall, this review of literature related to our research goals highlights the importance of clarity of the problem to be solved and collection of relevant data including proper labels for supervised machine learning. Additionally, adopting good pre-processing and feature selection strategies are central to all the implementations. Lastly, we found that the most compared algorithms for news topic classification are XGBoost, SVM, Naïve Bayes and Random Forest, and we intend to benchmark these in our experiments.

3.0 Dataset

Over the period of four months (October 2019 and January 2020), we scraped the websites of the British Broadcasting Corporation (BBC) and New York Times (NYT) for articles that include twitter URLs (Unique Resource Locators), irrespective of topic area. Our Python BeautifulSoup API based custom scrapper retrieved a total of 5,305 Record Tweets (RecTweets) from the news websites. This dataset is called the News-cited dataset. Using the TWARC utilities from the DocumentingNow project [25], the tweet IDs from the News-cited dataset were hydrated to JSONL format. Another python program was used in conjunction with the TWARC-Replies utility to retrieve over 2 million replies or Supporting Tweets (SupTweets) from the Twitter Search API stream.

Initial exploratory data analysis was performed on the data with the following outcome:

- Total 5,305 RecTweets extracted from BBC and NYT
    - 5,041 successfully hydrated from Twitter (264 could not be found - might have been deleted by the owners).
    - 4,708 were English language tweets.
    - 1,583 of the English language RecTweets were properly pre-classified into content categories by the news publishers.
- Total 2,429,549 SupTweets retrieved from Twitter Standard Search API[1] / TWARC-Replies
    - 1,980,781 were English language tweets
    - 4,451 of the 4,708 English language RecTweets had at least 1 SupTweet
    - Large variance observed in the number of SupTweets per RecTweets (from 0 to >40,000). See **Fig 1**.
- Total 14 topical categories were provided from the news publishers (see: **Table 1**). The categories will be used

as the class labels for the supervised machine learning models in this research.

[1] This research used the free Standard Search API, which provides free access to public tweets posted within the past 7 days only [26]. This restriction contributes to the unavailability of some SupTweets replies in our dataset.

Table 1: Topical Categories

ID	Class	All Content	SMR
1	Arts	14	11
2	Books	23	20
3	Business	139	119
4	Climate	12	12
5	Economy	12	12
6	Entertainment	464	429
7	Fashion	3	3
8	Food	12	12
9	Health	14	14
10	Politics	715	685
11	Sports	172	152
12	Technology	120	112
13	Travel	2	2
14	Unknown	3,603	-
		5,305	**1,583**

4.0 Pre-processing

The tweet text was cleaned up by removing all special characters, URLs, html codes, emojis, hashtags, and user mentions. The emojis, hashtags and user mentions were preserved in separate fields. Hashtags and mentions were de-duplicated to ensure uniqueness of the stored values for the cases where repeated in the RecTweets and SupTweets. All unicode characters were converted to ASCII using the *unidecode* Python module. For example, a word like Tánaiste was converted to Tanaiste. The NLTK module was used to tokenize the tweet text, perform stemming (PorterStemmer) and lemmatization (WordNet). We extended the standard NLTK stop-words list with an additional list of 290+ stop-words that are irrelevant to the classification task.

The pre-processing steps were done separately for RecTweets and SupTweets. All the replies (SupTweets) to a RecTweet were concatenated together to form a continues corpus of text. The RecTweet + SupTweets text is referred to as the Social Media Record (SMR) dataset in this paper and will be used for the NLP classification task.

SupTweets Threshold

To avoid overfitting and/or underfitting due to the large variance observed in the SupTweets, we limited the SMR dataset to SupTweets with minimum 10 replies and maximum 260 (corresponding to the median and upper whisker of the BoxPlot in **Fig 2**, respectively). Additionally, to ensure that the most relevant tweets were included in the SMR dataset, given the cap of 260 replies, we applied the following sort order to the SupTweet stream for each RecTweet:

- SupTweet size>=141, no URLs or Media, ordered by parsed_created_at
- SupTweet size<=140, no URLs or Media, ordered by parsed_created_at
- SupTweet with URLs and Media (potential spams)

We consider a greater than 140 characters sized SupTweet a large sized tweet given that it was the old limit set by Twitter [27] before the recent revision to 280 characters. This scheme allowed us to capture the more relevant supporting tweets before reaching the maximum threshold of 260 SupTweets.

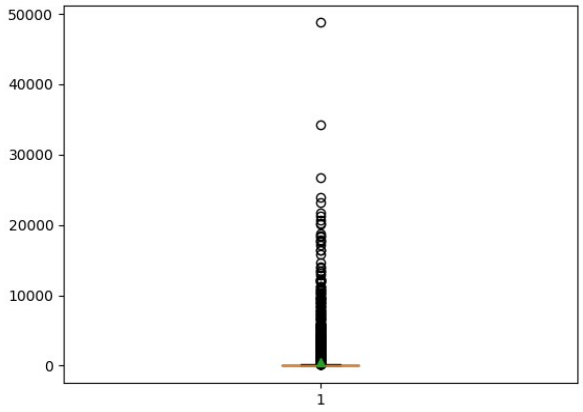

Fig 1: Supporting Tweets (Replies) with outliers

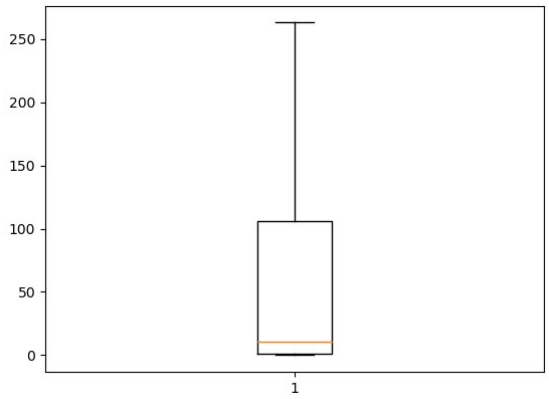

Fig 2: Supporting Tweets (Replies) without outliers

5.0 Experiments

Algorithms

Based on our literature review of related works, we selected the most benchmarked machine learning classifiers for our experiments. These are: Linear Support Vector Machine (LSVM), Gaussian Naïve Bayes (GNB), Random Forest (RF) and XGBoost.

Baseline Vectorizer

The pre-processed SMR dataset was converted to a unigram Document Text Matrix (DTM) using the Python sklearn CountVectorizer library as a baseline for the experiments. A preliminary review of the DTM revealed term sparsity issues. To reduce the sparsity, the following steps were taken:

A) Constant features removal

Constant features are terms in the DTM that are found in only a document (SMR). These constant features provide little or no value to the classification task [28] and can be removed from the DTM. Setting the minimum Document Frequency (*min_df*) parameter of the vectorizer to 2, accomplishes the removal of the constant features while creating the DTM.

B) Maximum features

It is good practise to set a maximum number of features for the vectorizer to use as a dimensionality reduction solution [29]. The number typical varies from 1,000 [30] to 20,000 [29] depending on the dataset. We set the baseline vectorizer's *max_features* parameter to 20,000. During the experiments, we vary the *max_features* parameter to 10,000 and 5,000 and observe performance metric changes to determine the optimum value for the parameter.

C) Training / Test Split

For the baseline machine learning experiment, we divided the SMR dataset into training and test dataset using a 80% and 20% respective split ratio [31]. We encountered some challenges with SMR text with low count of records in the labelled class as they do not divide well with the stratify parameter of the Vectorizer. To resolve the problem, we removed records where there are less than 20 items in a class from the SMR dataset. This reduced the classes available in the SMR dataset to 5 (see **Table 2**). We performed a 5-fold cross validation on the training data. We chose 5-fold cross validation due to the small size of the data [32].

Table 2: Final SMR dataset classes

ID	Class	SMR Count
3	Business	67
6	Entertainment	264
10	Politics	389
11	Sports	97
12	Technology	57
		874

Experimental results

We ran a total of 23 experiments with various vectorizer parameters and feature sets. We executed 9 baseline experiments using the Baseline Vectorizer described in the sub-section above and various elements of the SMR dataset. Experiment id "A8. Mentions+Nouns" produced the best baseline result with F1 scores ranging between 76% to 81% across all the 4 classifiers. The other baseline experiments produced metric scores ranging from 47% to 81%. We decided to carry forward the configuration for A8. as the baseline. For the simplicity of this report, we discuss the performance metrics in terms of F1 scores because it is considered the most comprehensive of the performance metrics [33] and derived from a combination of Precision and Recall, with the formula: F1 score = 2 * ((Precision * Recall) / (Precision + Recall)).

In the next series of experiments, we varied n-grams on the baseline and found that both "B1. 1-gram+2-gram" and "B2. 1-gram+2-gram+3-gram" performed well with F1 score ranges of 76% to 81% across all the 4 classifiers. Without a clear winner in this test, we kept the A8. Configuration, but replaced the Count Vectorizer with a TF-IDF Vectorizer. The TF-IDF test produced F1 score ranges of 76% to 82%, which is a slight improvement, but not enough to justify replacing the baseline for the next set of experiments.

Using the baseline Count Vectorizer, we changed the max_features parameter to experiment with 5,000 and 10,000 Top-N features. The results remained flat with F1 score ranges of 76% to 81%.

For the final set of tests, we replaced the Count Vectorizer with a TF-IDF Vectorizer and varied Top-N and n-Gram features. This time, we got a decisive improvement with the Linear SVM algorithm having an F1 score of 84.97% in the "F2. TFIDF 10k Features 1ng_2ng" test. The other performance measurement: accuracy, precision and recall scores for the Linear SVM (Mentions+Nouns, TF-IDF, 10,000 max_features, ngram=1,2) test were also above the other algorithms in the test group and the baseline test (see **Table 3**).

Table 3: Baseline vs Best result.

Experiment	Classifier	Exec_time	Accuracy	F1 score	Precision	Recall
A8. Mentions+Nouns	RandomForest	10.96	80.92	76.04	76.81	80.92
A8. Mentions+Nouns	LinearSVM	18.25	81.98	80.43	84.25	81.98
A8. Mentions+Nouns	GaussianNB	3.769	80.46	80.60	82.64	80.46
A8. Mentions+Nouns	XGBoost	219.2	83.97	81.69	85.61	83.97
F2. TFIDF 10k Features 1ng_2ng	RandomForest	11.97	82.12	77.17	78.78	82.12
F2. TFIDF 10k Features 1ng_2ng	**LinearSVM**	**39.82**	**86.84**	**84.98**	**87.95**	**86.84**
F2. TFIDF 10k Features 1ng_2ng	GaussianNB	4.353	79.40	79.16	81.30	79.40
F2. TFIDF 10k Features 1ng_2ng	XGBoost	252.2	83.83	81.55	85.23	83.83

6.0 Discussions and Conclusion

We had two primary goals set for our research. The first was to explore the properties of a social media records dataset for characteristics that can assist in automatic record selection and secondly, to use NLP techniques to determine the best parameters and algorithms for the training of a machine learning classifier. To achieve these goals, we collected data from two reliable news sources: the BBC and NYT and experimented with 4 machine learning algorithms to automate the selection of records into record classifications.

For the **first goal**, we found that user mentions and nouns (names of people, places, and things), when combined, are the best natural language properties of the News-cited SMR dataset for record selection automation. We also found

our scheme of assembling the social media record as a combination of the news-cited tweet and its replies, an effective lexical density strategy.

We fulfilled our **second goal** by performing 23 machine learning experiments using 4 different algorithms: Linear Support Vector Machine (LSVM), Gaussian Naïve Bayes (GNB), Random Forest (RF) and XGBoost. The results show that LSVM is the best performing algorithm with accuracy, F1, precision and recall scores of 86.84%, 84.98%, 87.95%, 86.84% respectively. While the SLVM program execution time of 39.87 seconds is not the fastest, it is much better than the slow running XGBoost time of 252.2 seconds.

Future work

There are several areas we would like to consider for a future work on social media records classification. We would like to experiment with more data to see if higher number of classes and features will improve the performance of the machine learning model. Also, we would like to use social media data from sources beyond news articles to broaden the record selection capability of the model. Additionally, it would be interesting to experiment with other feature engineering techniques such as Mutual Information Gain and advanced machine learning algorithms, for example, Deep Learning.

Limitations

Our research has certain limitations. One significant issue is media posting often include images, videos, and links to external sources. These are excluded from the dataset used in this research, since we focused on text-based NLP techniques. It is also important to note that the result of this research is only a framework. While our research illustrates how a social media records classification can be developed and utilized in records selection, it is not comprehensive, and is limited to the few categories discovered in the news-cited dataset.

Overall, despite the limitations of our research work, we have laid a foundation for the automation of social media records selection and records classification, both of which are essential for a meaningful records management approach to social media. The big data nature of social media especially in terms of unstructured natural language variety makes the use of machine learning compelling, if not a must-have solution. We consider our recommendations of the social media record structure (as a main record plus supporting replies), feature selection (mentions and nouns) and algorithm (linear SVM) worthy contributions towards a future, better managed social media ecosystem.

References

1. Walsh, Paul, Star Tribune June 11, 2020, https://www.startribune.com/teen-who-shot-video-of-george-floyd-wasn-t-looking-to-be-a-hero-her-lawyer-says/571192352, last accessed: 2020/09/11.
2. Pondiwa S., Phiri M., Challenges and Opportunities of Managing Social Media Generated Records in Institutions of Learning: A Case of the Midlands State University, Zimbabwe. In: Tatnall A., Mavengere N. (eds) Sustainable ICT, Education and Learning. SUZA 2019. IFIP Advances in Information and Communication Technology, vol 564. Springer, Cham (2019).
3. Zubiaga, Arkaitz. Mining social media for newsgathering: A review. Online Social Networks and Media 13 (2019).
4. Eggleston, E.M., Weitzman, E.R. Innovative Uses of Electronic Health Records and Social Media for Public Health Surveillance. Curr Diab Rep 14, 468 (2014).
5. Bertot, John. Social Media, Open Platforms, and Democracy: Transparency Enabler, Slayer of Democracy, Both? Proceedings of the 52nd Hawaii International Conference on System Sciences. (2019).
6. Franks, Patricia, and Alan Doyle. Retention and Disposition in the Cloud-Do You Really Have Control? Proceedings of The International Conference on Cloud Security Management ICCSM-2014. (2014).
7. NARA, National Archives and Records Administration. Guidance on managing social media records. https://www.archives.gov/records-mgmt/bulletins/2014/2014-02.html, last accessed: 2020/09/15.
8. Iron Mountain, Social Media Records call for fresh approach. https://www.ironmountain.com/resources/general-articles/s/social-media-records-call-for-fresh-approach, last accessed: 2020/09/17.
9. Cisco, Susan L., and Karen V. Strong. The value added information chain. Information Management 33.1, pp. 4. (1999).
10. Low, Jyue Tyan. A literature review: What exactly should we preserve? How scholars address this question and where is the gap. arXiv preprint arXiv:1112.1681 (2011).
11. Deacon, David. Yesterday's papers and today's technology: Digital newspaper archives and 'push button'content

analysis. European Journal of Communication 22.1. pp. 7. (2007).
12. Caron, Daniel, and Richard Brown. Appraising Content for Value in the New World: Establishing Expedient Documentary Presence. The American Archivist 76.1. 135-173. (2013).
13. Streck, Helen, and Endowment Fund. Social networks and their impact on records and information management. ARMA International Educational Foundation. 3-9. (2011).
14. K. Strutin, Social Media and the Vanishing Points of Ethical and Constitutional Boundaries, Pace Law Review, Volume 31, Issue 1, Article 6. 227-290. (2011).
15. Van Wyk, Hannah, and Kate Starbird. Analyzing Social Media Data to Understand How Disaster-Affected Individuals Adapt to Disaster-Related Telecommunications Disruptions. (2020).
16. Liu, Gaoyang, et al. SocInf: Membership inference attacks on social media health data with machine learning. IEEE Transactions on Computational Social Systems 6.5. 907-921. (2019).
17. Iman, Zahra, et al. A longitudinal study of topic classification on twitter. Eleventh International AAAI Conference on Web and Social Media. (2017).
18. Perreault, Mathieu, and Derek Ruths. The effect of mobile platforms on Twitter content generation. Fifth international AAAI conference on weblogs and social media. (2011).
19. Cohen, Raviv, and Derek Ruths. Classifying political orientation on Twitter: It's not easy! Seventh International AAAI Conference on Weblogs and Social Media. (2013).
20. Shahzad, Basit, et al. Discovery and classification of user interests on social media. Information Discovery and Delivery 45.3. 130-138. (2017).
21. Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. arXiv preprint arXiv:1610.07363 (2016).
22. Hassan, Noha Y., et al. Credibility Detection in Twitter Using Word N-gram Analysis and Supervised Machine Learning Techniques. International Journal of Intelligent Engineering and Systems 13.1. 291-300. (2020).
23. Helmstetter, Stefan, and Heiko Paulheim. Weakly supervised learning for fake news detection on Twitter. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, (2018).
24. Salminen, Joni, et al. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. ICWSM. (2020).
25. TWARC, Documenting The Now, https://github.com/DocNow/twarc, last accessed: 2020/09/26.
26. Twitter, Standard Search API, https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview/standard, last accessed: 2020/09/26.
27. Gligorić, Kristina, Ashton Anderson, and Robert West. How constraints affect content: The case of Twitter's switch from 140 to 280 characters. arXiv preprint arXiv:1804.02318. (2018).
28. Chouhan, Ashish, and Ajinkya Prabhune. FIF: A NLP-based Feature Identification Framework for Data Warehouses. 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI). IEEE, (2019).
29. Gimpel, Kevin, et al. Part-of-speech tagging for twitter: Annotation, features, and experiments. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, (2010).
30. Curiskis, Stephan A., et al. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management 57.2. 102034. (2020).
31. Abro, Sindhu, et al. Aspect Based Sentimental Analysis of Hotel Reviews: A Comparative Study. Sukkur IBA Journal of Computing and Mathematical Sciences 4. 11-20. (2020).
32. Hollenstein, Nora, et al. Advancing NLP with cognitive language processing signals. arXiv preprint arXiv:1904.02682. (2019).
33. Lee, Lung-Hao, Liang-Chih Yu, and Li-Ping Chang. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications. (2015).